文書データの符号処理方法及びシステム

(METHOD AND SYSTEM FOR CODE PROCESSING OF DOCUMENT DATA)

発明の技術分野(FIELD OF THE INVENTION)

本発明は、文書データの符号処理方法及びシステムに関する。

関連技術の説明(DESCRIPTION OF THE RELATED ART)

従来、伝送すべきデータ量を削減するために、文書データを符号化及び復号化する方法がある。この方法を実現するには、送信装置及び受信装置はそれぞれ、変換テーブルを所持する必要がある。変換テーブルは、記述言語と符号データとを１対１に対応付けたものである。送信装置は、変換テーブルに基づいて文書データを符号データに符号化する。一方、受信装置は、変換テーブルに基づいて符号データを文書データに復号化する。

このような方法は、特に、インターネットに有効である。例えば、Webサーバが、HTML(HyperText Markup Language)のようなテキスト形式のマーク付け言語で記載された文書データを符号化した符号データを送信する。これに対し、クライアントが、受信した符号データを文書データに復号化し、その文書データをブラウザに表示する。このとき、文書データを符号データで伝送することにより、伝送データ量を削減することができる。

インターネットにおいて、文書データを符号化することは、セキュリティの観点からも有効である。変換テーブルを有さないクライアントは、符号データを復号することができないからである。

図１は、従来の文書データの符号化及び復号化方法である。図１によれば、HTML形式の文書データ１２は、変換テーブル１１を用いて符号化１０される。一方、符号データは、変換テーブル２１を用いて復号化２０される。これにより、HTML形式の文書データ２２が復元される。更に、文書データ２２は、パーサ２３によって要素の論理構造が解析され、ブラウザ２４を用いて表示される。図１によれば、符号化で用いられる変換テーブル１１と、復号化で用いられる変換テーブル２１は、同一テーブルである必要がある。

しかしながら、近年では、Webサーバが送信する文書データは、HTMLのように情報の表示を規定するだけでなく、情報の構造を規定することもできる拡張可能なテキスト形式のマーク付け言語が多くなってきた。この言語は、例えば、XML(eXtensible Markup Language)又はSGML(Standard Generalized Markup Language)である。

例えば、図１のような従来の符号化及び復号化方法によれば、文書データが拡張されると、変換テーブルも拡張しなければならない。

また、マーク付け言語は、要素の論理構造を規定しているので、必ず、符号データを文書データに復号し、その文書データの論理構造をパーサによって解析し且つ処理する必要がある。

発明の要約(SUMMARY OF THE INVENTION)

本発明の目的は、拡張可能なテキスト形式の記述言語を符号化することができ、符号データを文書データに復号することなく文書処理をすることができる文書データ符号処理方法及びシステムを提供することである。

本発明の文書データの符号処理方法によれば、拡張可能なテキスト形式の記述言語で記載された変換テーブルを用いて、拡張可能なテキスト形式の記述言語で記載された文書データを符号データに符号化する符号化ステップと、変換テーブルを用いて、符号データを文書データとして文書処理する文書処理ステップとを有し、変換テーブルは、他の変換テーブルのリンク情報を定義し、要素名と、該要素名の要素値と、該要素名に指定可能な属性名と、該属性名の属性値との項目に割り当てられた符号長及び符号と、第１の要素名に対する第２の要素名が親子関係にある論理構造を示す符号長及び符号とを定義する。

これにより、変換テーブル自体が拡張可能であるので、拡張可能な文書データに対応することができる。また、変換テーブルによって論理構造を符号データに含ませることができるので、文書データに復号し且つパーシングすることなく、直接的に文書処理を行うことができる。このような効果は、例えば携帯電話機のように、低い処理能力しか有さない受信装置にとって、処理負荷が小さいという効果を奏する。

本発明の他の実施形態によれば、文書処理ステップで用いる変換テーブルに定義されている項目は、符号化ステップで用いる変換テーブルに定義されている項目の部分集合の関係にあることも好ましい。

例えば、一方の受信装置では、変換テーブルのある部分のみが所持され、他方の受信装置では、変換テーブルの他の部分のみが所持されているとする。そして、送信装置は、１つの文書データを符号化した符号データを、複数の受信装置へ配信する。この結果、一方の受信装置では、文書データのある部分のみが表示され、他方の受信装置では、文書データの他の部分のみを表示されるようになる。配信する符号データは同じであっても、受信装置が所持する変換テーブルによって文書処理された表示が異なることになる。このような機能は、セキュリティの観点からも有効なものである。

本発明の他の実施形態によれば、符号化ステップは、文書データに、変換テーブルに定義されていない他の変換テーブル、要素名、要素値、属性名及び属性値が存在する場合、該要素名、要素値、属性名及び属性値を符号化しないことも好ましい。

これにより、文書データの一部が符号化できないという理由で、文書データ全体が符号化できないとすることを避けることができる。

本発明の他の実施形態によれば、符号化ステップは、変換テーブルを用いて符号化した部分の占有データ長を該変換テーブルを指示する符号に付加し、項目それぞれが占有する部分の占有データ長をそれぞれの項目の符号に付加し、文書処理ステップは、変換テーブルに定義されていない符号が、符号データに存在する場合、該符号を文書処理せずに、占有データ長を飛ばした位置の符号データから文書処理することも好ましい。

これにより、符号データにおける文書処理をしない部分を読み飛ばすことができる。

本発明の文書データの符号処理システムによれば、拡張可能なテキスト形式の記述言語で記載された文書データを送信するサーバと、受信した文書データを、変換テーブルを用いて符号データに符号化する符号化サーバと、受信した符号データを、変換テーブルを用いて文書処理する手段を有するクライアントと

を有し、変換テーブルは、拡張可能なテキスト形式の記述言語で記載されており、他の変換テーブルのリンク情報を定義し、要素名と、該要素名の要素値と、該要素名に指定可能な属性名と、該属性名の属性値とに割り当てられた符号長及び符号と、第１の要素名に対する第２の要素名が親子関係にある論理構造を示す符号長及び符号とを定義する。

これにより、既存のサーバをそのまま利用することができる。

本発明の他の実施形態によれば、符号化サーバで用いる変換テーブルに定義されている項目は、クライアントで用いる変換テーブルに定義されている項目の部分集合の関係にあることも好ましい。

本発明の他の実施形態によれば、符号化サーバは、文書データに、変換テーブルに定義されていない他の変換テーブル、要素名、要素値、属性名及び属性値が存在する場合、該要素名、要素値、属性名及び属性値を符号化しないことも好ましい。

本発明の他の実施形態によれば、符号化サーバは、変換テーブルを用いて符号化した部分の占有データ長を該変換テーブルを指示する符号に付加し、項目それぞれが占有する部分の占有データ長をそれぞれの項目の符号に付加し、クライアントは、変換テーブルに定義されていない符号が、符号データに存在する場合、該符号を文書処理せずに、占有データ長を飛ばした位置の符号データから文書処理することも好ましい。

本発明の更なる目的及び効果は、添付図面に表された、本発明の好ましい実施形態の以下の説明から明らかとなる。

図面の簡単な説明(BRIEF DESCRIPTION OF THE DRAWINGS)

図１は、従来の基本的な符号化及び復号化方法の説明図である。

図２は、本発明による符号文書処理方法の説明図である。

図３は、ＸＭＬ形式の文書データのサンプルである。

図４は、図３の文書データの符号データの一例である。

図５ａは、図３の文書データを、図４の符号データに変換するための変換テーブルである。特にヘッダ部分のテーブルである。

図５ｂは、図３の文書データを、図４の符号データに変換するための変換テーブルである。特にルート要素のテーブルである。

図５ｃは、図３の文書データを、図４の符号データに変換するための変換テーブルである。特に第１の子要素のテーブルである。

図５ｄは、図３の文書データを、図４の符号データに変換するための変換テーブルである。特に第２の子要素のテーブルである。

図６は、他の変換テーブルのリンク情報を含む変換テーブルである。

図７は、要素毎に該要素の占有する占有データ長を付加した符号データである。

図８は、本発明の第１の実施形態のシステム構成図である。

図９は、本発明の第２の実施形態のシステム構成図である。

図１０は、本発明の文書処理のフローチャートである。

好ましい実施形態の説明(DESCRIPTION OF THE PREFERRED EMBODIMENTS)

図２は、本発明による文書データの符号処理方法である。図２によれば、文書データ１２は、複数の文書データ１２０及び１２１によって拡張されている。一方、変換テーブル１１も、拡張された文書データに対応して、複数の変換テーブル１１０及び１１１のリンク情報を定義している。これにより、XML形式の文書データ１２は、変換テーブル１１を用いて符号化１０される。

また、図２によれば、符号データは、変換テーブル２１を用いて、直接的に文書処理３０され、ブラウザ２４に表示される。本発明によれば、符号データには、要素の論理構造も含まれる。従って、文書データに復号する必要もなく、更にパーサ２３によって論理構造を解析する必要もない。

図３は、文書データのサンプルである。図４は、図３の文書データを符号化した符号データである。図５ａ～ｄは、図３の文書データの変換テーブルである。以下では、図３及び図４を参照しつつ、図５ａ～ｄの変換テーブルを説明する。

変換テーブルは、XMLで記載されており、図５ａのヘッダ部分<head>(1)と、図５ｂ～ｃのボディ部分<body>(8)とに分けられる。ヘッダ部

分には、接頭辞について記述する。ボディ部分には、文書の論理構造と変換符号とを記述をする。

図５ａによれば、ヘッダ部分には、接頭辞の符号長(2)として2bitが割り当てられる。接頭辞として、要素名及び属性名には符号"00"(3)が割り当てられる。また、要素値及び属性値の内容が、数値であれば符号"01"(4)が、文字列であれば符号"10"(5)が割り当てられる。

更に、図３には、要素名SVGが定義されているので、図５ａによれば、要素名"SVG"の開始に3bit"000"(6)が割り当てられ、その終了に3bit"011"(7)が割り当てられる。

図５ｂによれば、最初に要素名SVGを以下で定義する(9)ことを表している。また、この要素名SVGに付随する属性名に、2bitの符号長を割り当てる(10)ことを定義する。そして、属性名widthに符号"10"を割り当て(11)、属性名heightに符号"11"を割り当てる(13)。また、属性名widthの属性値は、符号無し整数10bitで表され(12)、属性名heightの属性値も、符号無し整数10bitで表される(14)。

次に、要素名SVGの子関係にある要素を符号長3bitで表す(15)ことを定義する。そして、要素名SVGの子要素として要素名rectを定義する(16)。要素名rectの開始に符号"001"が割り当てられ、その終了に符号"011"が割り当られる(17)。また、要素名SVGの子要素として要素名textを定義する(18)。要素名textの開始に符号"010"が割り当てられ、その終了に符号"011"が割り当てられる(19)。

図５ｃによれば、次に、要素名rectを以下で定義する(20)ことを表している。また、この要素名rectに付随する属性名に、3bitの符号長を割り当てる(21)ことを定義する。属性名xには符号"100"を割り当て(22)、属性名xの属性値は符号付き整数10bitで表される(23)。また、属性名yには符号"101"を割り当て(24)、属性名yの属性値は符号付き整数10bitで表される(25)。また、属性名widthは符号"110"を割り当て(26)、属性名widthの属性値は符号無し整数10bitで表される(27)。最後に、属性名heightには符号"111"を割り当て(28)、属性名widthの属性値は符号無し整数10bitで表さ

れる(29)。

図５ｄによれば、次に、要素名 text を以下で定義する(30)ことを表している。また、この要素名 text に付随する属性名に、2bit の符号長を割り当てる(31)ことを定義する。属性名 x には符号"10"を割り当て(32)、属性名 x の属性値は符号付き整数 10bit で表される(33)。また、属性名 y には符号"11"を割り当て(34)、属性名 y の属性値は符号付き整数 10bit で表される(35)。

次に、要素 text の要素値を以下で定義する(36)を表している。ここでは、要素値が Shift-JIS 形式であることを表している(37)。

図６は、複数の変換テーブルのリンク情報を定義した変換テーブルの一例である。本発明が対象としている記述言語は、拡張可能なテキスト形式の記述言語である。従って、文書データが拡張されると同様に、変換テーブルも拡張する必要がある。図６によれば、ヘッダ部分に複数の変換テーブルのリンク情報を定義するだけで、変換テーブルを作成し直す必要がない。ヘッダ部分には、複数の変換テーブルを拡張するためのメタ情報を定義する。メタ情報とは、接頭辞符号の符号・符号長、要素の指定、名前空間の指定、変換テーブルへのリンク情報である。

図７は、要素が占有する占有データ長を、図４の符号データに埋め込んだものである。これにより、文書処理を行うクライアントは、所持する変換テーブルに定義されていない符号が符号データに存在する場合、該符号を文書処理せずに、占有データ長を飛ばした位置の符号データから文書処理することができる。

図８は、本発明の第１の実施形態のシステム構成図である。図８によれば、サーバ４は、予めクライアントＡ及びＢへ、変換テーブルを送信する。この場合、サーバ４の所持する変換テーブルの項目の部分集合となる変換テーブルａ及びｂをそれぞれ送信する。その後、サーバ４は、クライアントＡ及びＢへ、文書データを符号化した符号データを送信する。この符号データを受信したクライアントＡ及びＢはそれぞれ、文書処理を行うが、実際にブラウザに表示される情報は、異なるものとすることができる。

図９は、符号化サーバ６を含むシステム構成図である。サーバ４は、ＸＭＬ形式の文書データを符号化サーバ６へ送信する。符号化サーバ６は、変換テーブルサーバ７から受信した変換テーブルを用いて、文書データを符号化する。その符号データは、クライアント５へ送信される。クライアント５は、変換テーブルサーバ７から受信した変換テーブルを用いて、文書処理を行う。図９によれば、ＸＭＬ形式の文書データを送信する既存のサーバに変更を加えることなく、符号化サーバをプロキシサーバとして利用することができる。

図１０は、文書処理のフローチャートである。例えば、図４の符号データを、図５の変換テーブルに基づいて行う文書処理を説明する。

（Ｓ１）変換テーブル<head><prefix bit="2">によればヘッダ符号長 2bit であるので、符号データから 2bit を読み込む。図４によれば"00"であるので、"名"を示す符号であると判断する。

（Ｓ２）次に、変換テーブル<head><root name="svg" bit="3" code="000" />によればルート要素"svg"であり、次の 3bit を読み込む。符号は"000"であるので、要素 svg の開始であると解釈する。

（Ｓ３）符号データからヘッダ符号長 2bit を読み込む。

（Ｓ４）図４によれば"00"であり、変換テーブルの<head>から"00"は"名"を示す符号であると判断する。

（Ｓ５）属性名の符号長<attlist bit=2>、子要素名の符号長<children bit=3>、終了タグ<end name="/svg" bit=3 code="011"/>の中で、最も短い符号長分 2bit だけ読み込む。

（Ｓ６）図４によれば"10"であるので、属性名 width と一致することを確認する。

（Ｓ７）もし、一致しなかった場合、次に短い符号長分 3bit を読み込み、再びＳ６へ戻る。

（Ｓ８）"10"と一致しているので、属性名 width と解釈する。

（Ｓ９）次の 3bit が、終了タグ<end name="/svg" bit=3 code="011"/>でないことを確認する。終了タグであれば、終了する。終了タグでなければ、再びＳ３へ戻る。

（Ｓ３）符号データからヘッダの符号長"2"bitを読み込む。

（Ｓ４）図４によれば"01"であり、変換テーブル<head><number_prefix code="01" />によれば"01"は"数値"を示す符号であると判断する。

（Ｓ１０）変換テーブル<number bit="10" data="UI" qt="1" />によれば、属性名widthの属性値は符号無し整数10bitであるので、10bitを読み込む。

（Ｓ１１）"0111110100"であるので、属性値"500"と解釈する。そして、再びＳ３へ戻る。

前述したように図１０のフローチャートを繰り返すことによって、符号データを復号することなく直接的に文書処理をすることが可能となる。

以上、詳細に説明したように、本発明によれば、拡張可能なテキスト形式の記述言語によって記載された文書データの符号化を可能とする。このような符号化は、データ伝送量を削減することができるので、無線のような伝送速度が遅い通信システムに効果がある。

また、本発明によれば、拡張可能なテキスト形式で記述された文書データに対して、符号化装置を変更することなく、変換テーブルを置き換えるだけで、それぞれの文書データに適した符号化を行うことが可能となる。更に、文書データが拡張された場合でも、元の文書データ用の符号化テーブルは変更せず、拡張部分のみの符号化テーブルを用意するだけで、拡張された文書データに適した符号化を行うことができる。

また、本発明によれば、復号側装置に文書専用処理エンジンを搭載することにより、受信した符号データから元の文書データを復元させる必要がなく、復号側装置にとって処理負荷が小さいという効果がある。

本発明の多くの広範な種々の実施形態は、本発明の技術的思想及び見地から離れることなく構成されている。本発明は、

以上述べた実施形態は全て本発明を例示的に示すものであって限定的に示すものではなく、本発明は他の種々の変形態様及び変更態様で実施することができる。従って本発明の範囲は特許請求の範囲及びその均等範囲によってのみ規

定されるものである。